



Structural Diversity and Homophily: A Study Across More than One Hundred Large-Scale Networks

Yuxiao Dong, Reid A. Johnson, Jian Xu, Nitesh V. Chawla
 Department of Computer Science and Engineering
 Interdisciplinary Center for Network Science and Applications (iCeNSA)
 University of Notre Dame, Notre Dame, IN 46556
 {ydong1, rjohns15, jxu5, nchawla}@nd.edu

ABSTRACT

Understanding the ways in which local network structures are formed and organized is a fundamental problem in network science. A widely recognized organizing principle is structural homophily, which suggests that people with more common neighbors are more likely to connect with each other. However, what influence the diverse structures formed by common neighbors (e.g.,  and ) have on link formation is much less well understood. To explore this problem, we begin by formally defining the structural diversity of common neighborhoods. Using a collection of 116 large-scale networks—the biggest with over 60 million nodes and 1.8 billion edges—we then leverage this definition to develop a unique network signature, which we use to uncover several distinct network superfamilies not discoverable by conventional methods. We demonstrate that structural diversity has a significant impact on link existence, and we discover striking cases where it violates the principle of homophily. Our findings suggest that structural diversity is an intrinsic network property, giving rise to potential advances in the pursuit of theories of link formation and network evolution.

Keywords

Social Diversity; Homophily; Network Motif; Triadic Closure; Network Superfamily; Link Prediction; Big Data.

1. INTRODUCTION

Since the time of Aristotle, it's been known that people “love those who are like themselves” [1]. We now know this as the principle of homophily, which suggests that the tendency of individuals to associate and bond with similar others drives the formation of social relationships [21, 28]. The powerful effects of homophily pervade our everyday lives, silently influencing our most basic relationships from friendship to marriage [28]. By guiding the formation of relationships, homophily also plays an important role in the dissemination of information, behavior, and even health [6]. But homophily applies to more than shared traits or characteristics: it applies to the fundamental structure of our relationships as well.

Structural homophily holds that individuals with more friends in common are more likely to associate [32, 18, 17]. The tendency of individuals to connect based on structural homophily has been widely explored in network science [25, 17, 3, 9], where it has been shown to be a strong driving force of link formation over a large assortment of networks. Yet, while structural homophily accounts for similarity based on the actual number of common neighbors, it fails to account for the diverse ways in which these neighbors may be connected—a phenomenon known as structural diversity. Despite the importance of structural homophily, the effects of structural diversity are much less well understood. This leaves many interest-

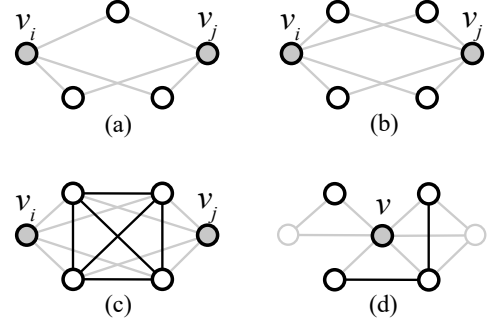




Figure 1: Structural diversity of common neighborhoods. (a) Two nodes v_i and v_j with three disconnected common neighbors. (b) v_i and v_j with four disconnected common neighbors. (c) v_i and v_j with four connected common neighbors. (d) Node v 's ego network with four out of six neighbors highlighted.

ing questions concerning the role of structural diversity in common neighborhoods unanswered, including: how it manifests across networks, how it varies according to the type of network, how well it concords with the principle of homophily, and how much it influences network connectivity in a neighborhood and beyond.

Motivating example. Consider the real-world scenarios presented in Figure 1. According to structural homophily, the probability that two users v_i and v_j know each other given that they share four common neighbors (CN), as shown in Figure 1(b), is generally higher than when they share only three, as shown in Figure 1(a). Formally, $P(e_{ij}=1 \mid \#CN=4) > P(e_{ij}=1 \mid \#CN=3)$, where $e_{ij}=1$ denotes the existence of an edge e between v_i and v_j . A natural question that arises is how two users' common neighborhood—that is, the subgraph structure of their common neighbors—influences the probability that they form a link in the network. For example, let us assume that v_i and v_j share four common neighbors. Are v_i and v_j more likely to connect with each other if their four common neighbors do not know each other (i.e., ) , as in Figure 1(b), or if they all know each other (i.e., ) , as in Figure 1(c)? In essence, then, we are interested in the truth of the following inequality:

$$P(e_{ij}=1 \mid \text{disconnected}) \gtrless P(e_{ij}=1 \mid \text{connected}) ?$$

In this work, we formally define the structural diversity of common neighborhoods between two individuals and study its impact on the probability that these individuals know each other. Our definition of the structural diversity of a common neighborhood is a mixture of the variety and density of common social contexts. Variety measures the number of connected components that comprise the common neighborhood, capturing the variable ways in which a neighborhood may be composed. Density measures the ratio be-

tween the number of edges and all possible edges among common neighbors, capturing how tightly connected the neighborhood is. In general, we consider a common neighborhood with more components and lower density to be more structurally diverse. Using this formulation, we study the influence of structural diversity on the formation of (social) relationships across more than one hundred large-scale networks from a wide range of domains (cf. Section 2), *making it the largest empirical analysis done on networks to date.*

The measure of structural diversity was first proposed in a study by Ugander et al. [39], which found that the user recruitment rate in Facebook is determined by the variety of an individual’s contact neighborhood. They focused on an ego-centric notion of structural diversity. We go beyond an ego, and focus on the *structural diversity of common neighborhoods* for two-person ego networks (see examples in Figure 1(a)(b)(c)), rather than the diversity of a sole individual’s ego network (see v ’s ego network in Figure 1(d)).

We leverage our definition of the structural diversity of common neighborhoods to develop a signature for network superfamily detection. By employing the structural diversity signature of each network, we are able to cluster all the real-world large-scale networks used in our study into three distinct superfamilies, each of which displays unique link existence patterns. We find that, surprisingly, these superfamilies cannot be uncovered by subgraph significance profiles [29], indicating that the structural diversity signature provides the ability to unveil previously undiscovered mechanisms of network organization. This finding demonstrates that the structural diversity of common neighborhoods can effectively capture driving forces intrinsic to the formation of local network structures.

We examine how the structural diversity of common neighborhoods impacts link existence and network connectivity. We also provide in-depth investigations using three large-scale networks—Friendster, BlogCatalog, and YouTube—each of which represents a particular network superfamily.¹ These findings generally hold for all networks within a given superfamily. Figure 2 reports the relative link existence rate between each pair of users who have at least one common neighbor, conditioned on several representative common neighborhoods (x -axis). We observe that with the same edge density and similar number of components, the link existence rates given different common neighborhoods (e.g., $\begin{smallmatrix} \circ & \circ \\ \diagup & \diagdown \end{smallmatrix}$, $\begin{smallmatrix} \circ & \circ \\ \diagup & \diagup \end{smallmatrix}$, and $\begin{smallmatrix} \circ & \circ \\ \diagdown & \diagdown \end{smallmatrix}$) are close to each other. In Friendster and its network superfamily, we find that when we fix the size of the common neighborhood (e.g., three common neighbors), an increase in the structural diversity of the neighborhood (i.e., greater variety and lower density) negatively impacts the formation of online friendships—that is, $P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagdown \end{smallmatrix}) < P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagup \end{smallmatrix})$. By contrast, in BlogCatalog and its superfamily, we discover that an increase in the structural diversity of the common neighborhood actually facilitates link formation—that is, $P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagdown \end{smallmatrix}) > P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagup \end{smallmatrix})$. We further note that other network properties show strong similarities between different superfamilies, and thus cannot adequately characterize the differences we observe in the structural diversity of common neighborhoods.

We also discover striking phenomena where structural diversity violates the principle of homophily. Formally, when applied to the context of common neighborhoods, the principle suggests that $P(e=1|\text{\#CN}=4) > P(e=1|\text{\#CN}=3)$. However, if we consider BlogCatalog, for example, we find that the link existence rate of four common neighbors in a single component is significantly lower than the rate of only three disconnected common neighbors, i.e., $P(e=1|\begin{smallmatrix} \circ & \circ & \circ \\ \diagup & \diagdown & \diagdown \end{smallmatrix}) < P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagdown \end{smallmatrix})$. Similarly, in Friendster, homophily is violated when comparing four disconnected common neighbors with three connected common neighbors, i.e.,

¹For full results over all of these networks, please refer to the companion webpage at <http://www3.nd.edu/~ydong1/diversity/diversity.html>.

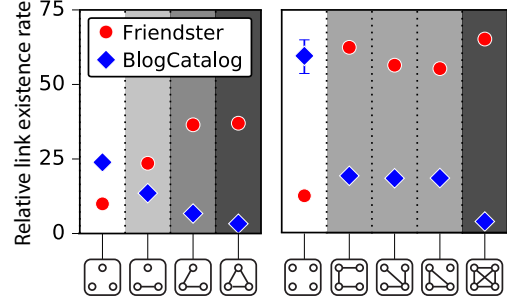


Figure 2: Structural diversity of common neighborhoods affects link existence rate. x -axis: all three-node common neighborhoods (left) and five out of eleven four-node common neighborhoods (right). y -axis: relative link existence rate between two users (cf. Section 3 for detailed definition).

$P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagdown \end{smallmatrix}) > P(e=1|\begin{smallmatrix} \circ & \circ \\ \diagup & \diagup \end{smallmatrix})$. These findings imply that structural diversity is a simple yet effective predictor for inferring link existence. For example, by applying structural diversity to the link inference task, we find a 57% improvement over structural homophily as measured by AUPR in both the Friendster and BlogCatalog Networks.

Studying structural diversity in the context of common neighborhoods sheds light on the pursuit to truly understand the driving forces behind the organization of neighborhoods in social networks. Our findings also have important, practical implications for recommendation functions in social networks, such as “People You May Know” in Facebook and “Who to Follow” in Twitter.

Organization. The remainder of the paper is organized as follows. In Section 2, we present the extensive set of large-scale network datasets used in this work. In Section 3, we introduce our definition of the structural diversity of common neighborhoods, noting several questions that this definition raises. In Section 4, we apply this definition to catalog our set of over one hundred large-scale networks into network superfamilies based on their structural diversity. In Section 5, we use representative networks from these superfamilies to examine how the structural diversity of common neighborhoods impacts link existence. Finally, we review related work in Section 6, and provide our conclusions in Section 7.

2. BIG NETWORK DATA

To comprehensively examine our proposed concept of the structural diversity of common neighborhoods, we have assembled a large collection of big network datasets from several well-known data platforms, including (in alphabetical order): AMiner (AMiner Open Science Platform [37])², ASU (Social Computing Data Repository [44])³, KONECT (Koblenz Network Collection [20])⁴, MPI (Social Computing Research at MPI-SWS [30])⁵, ND (Notre Dame [4])⁶, NetRep (Network Data Repository [34])⁷, Newman [33]⁸, and SNAP (Stanford Large Network Dataset Collection [24])⁹.

²<https://aminer.org/>

³<http://socialcomputing.asu.edu/pages/datasets>

⁴<http://konect.uni-koblenz.de>

⁵<http://socialnetworks.mpi-sws.org/datasets.html>

⁶<http://www3.nd.edu/~networks/resources.htm>

⁷<http://www.networkrepository.com/>

⁸<http://www-personal.umich.edu/~mejn/netdata/>

⁹<http://snap.stanford.edu/data/index.html>

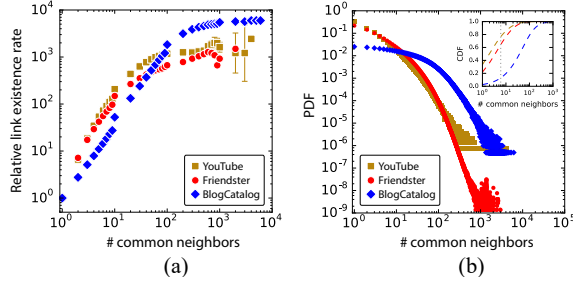


Figure 3: Common neighbor characterization. (a) The link existence rate as the function of #common neighbors. (b) The probability density function (PDF) of #common neighbors.

In total, we have compiled a set of 116 large-scale undirected and unweighted networks from the platforms listed above, including 84 real-world networks and 32 random graphs. We have cleaned the networks as follows: For directed social networks, such as mobile phone and SMS networks, we retain only reciprocal connections as undirected edges. For other directed networks that have no reciprocal connection, such as citation networks, we convert each directed link into an undirected one. We have then pruned the resulting undirected networks by removing all duplicate edges and self-loops, retaining only the largest connected component.

The detailed information about the 116 networks is shown in Table 4 (see the last page), which provides detailed network characteristics, including: order and size; the average, 10%, 50%, 90%, and maximum degrees; the average and global clustering coefficients (cc); diameter; and the number of closed triangles and triplets. Due to the large set of networks, we have labeled them according to the follow nomenclature: *type-original-platform*. *type* denotes the network type, most of which have been previously designated by their source data platforms (see the taxonomy example in SNAP). For a given network, *type* can be one of social blog-based (*blog-*), collaboration (*ca-*), citation (*cit-*), communication (*comm-*), location-based (*loc-*), or online social (*soc-*) networks, or web hyperlink graphs (*web-*). *original* denotes the original name of the networks as provided by each platform. *platform* denotes the data platform from which the network has been sourced. We note that the largest network used in this work is the *soc-Friendster-SNAP* online social network, which consists of over 65 million nodes and 1.8 billion edges.

To ensure that our study explores a representative sample of network structures, we also test the concept of structural diversity of common neighborhoods on 32 random graphs, including networks generated by the Erdős-Rényi (ER) model [13], Barabási-Albert (BA) model [4], and Watts-Strogatz (WS) model [43]. For all three models, the number of nodes is set as 1,000,000. In the ER model, we set the edge creation probability to between 5×10^{-6} and 5×10^{-5} with a step of 5×10^{-6} , thereby generating ten ER random graphs with the number of edges ranging from roughly 2,000,000 to 25,000,000. We use the BA model to generate eight BA random graphs with between 2,000,000 and 16,000,000 edges. Finally, we use the WS model to generate 14 WS random graphs by setting different mean degrees k and rewiring probabilities β , where k is chosen from 4, 8, 12, 16, 20, 24, and 28, and β is 0.2 or 0.8. There are between 2,000,000 and 14,000,000 edges in the WS graphs.

3. STRUCTURAL DIVERSITY OF COMMON NEIGHBORHOODS

In this section, we investigate the principles that drive the formation and organization of local structures in large-scale networks.

In particular, we focus on the common neighborhood of each pair of users and ask the following question: How does a pair of individuals' common neighborhood—the subgraph with their mutual neighbors as nodes and the connections among them as edges—influence the probability that there exists a link between them?

Formally, we use $G = \{V, E\}$ to denote an undirected and unweighted network, where $V = \{v_i\}$ represents the set of nodes and $E \subseteq V \times V$ represents the set of links between two nodes. We denote each existing link, $e_{ij} \in E$, as $e_{ij} = (v_i, v_j) = 1$ and each non-existing link, $e_{ij} \notin E$, as $e_{ij} = (v_i, v_j) = 0$.


Definition 1. Common Neighborhood: Let $N(v_i)$ denote the adjacency list of a node v_i , i.e., v_i 's neighborhood. The common neighborhood of each pair of two nodes v_i and v_j can be represented as the subgraph composed of their common neighbors, $G^{ij} = \{V^{ij}, E^{ij}\}$, where $V^{ij} = N(v_i) \cap N(v_j)$ denotes the common neighbors of v_i and v_j , and $E^{ij} = \{e_{pq} \mid e_{pq} \in E, v_p \in V^{ij}, v_q \in V^{ij}\}$ denotes the edges among their common neighbors.

Input: Given a network $G = \{V, E\}$, the input of our problem includes 1) each pair of users who have at least one common neighbor, i.e., $\{(v_i, v_j) = e_{ij} \mid |V^{ij}| \geq 1\}$, and the common neighborhood $G^{ij} = \{V^{ij}, E^{ij}\}$ of each pair of users v_i and v_j .

Structural homophily. The principle of structural homophily suggests that with more common neighbors, it is more likely for two people to know each other. Formally, this means that if $y > x$, then $P(e_{ij}=1 \mid |V^{ij}|=y)$ should generally be larger than $P(e_{ij}=1 \mid |V^{ij}|=x)$. A long line of work from various fields has demonstrated that this principle holds across a wide variety of different networks. For example, Figure 3(a) reports the link existence rate between two users (y -axis), conditioned on the size of their common neighborhood (x -axis) in three representative networks. We can see that as the number of common neighbors increases, the probability that two users are connected with each other increases in all three networks as well.

In this study, we revisit this principle of structural homophily, further proposing the concept of structural diversity of common neighborhoods. We define the structural diversity of a graph as a function of its variety and density, as formalize its application to common neighborhoods below.

Definition 2. Structural Diversity of Common Neighborhoods: Given a network, G , a pair of users in this network, v_i and v_j , and the pair's common neighborhood, $G^{ij} = \{V^{ij}, E^{ij}\}$, we define the structural diversity of G^{ij} as a mixture of its variety, $|C(G^{ij})|$, and density, $d(G^{ij})$, where $C(G^{ij})$ denotes the connected components in G^{ij} and $d(G^{ij})$ denotes the density of G^{ij} .

Consider a pair of users with four common neighbors. This pair's common neighborhood has 11 possible configuration structures: . In general, we refer to the more diverse structure as the one with fewer edges and more components.

Output: Our goal is to study the relations between the structure of two users' common neighborhood and the probability that there exists a link between these two users. Therefore, given two users v_i and v_j and their common neighborhood G^{ij} , the output of our problem is the link existence probability distribution of G^{ij} , $P(e_{ij}=1 \mid G^{ij})$.

In each network, we enumerate all pairs of users who have at least one common neighbor. If we fix the common neighborhood G^{ij} of two users v_i and v_j , then we can compute the link existence probability $P(e_{ij}=1 \mid G^{ij})$ based on the number of link pairs that

exist. To facilitate the comparability of results across networks with diverse sizes and densities, we define the relative link existence rate $R(e_{ij}=1 \mid G^{ij})$ as

$$R(e_{ij}=1 \mid G^{ij}) = \frac{P(e_{ij}=1 \mid G^{ij})}{P(e=1 \mid \#CN=1)},$$

where $P(e=1 \mid \#CN=1)$ denotes the link existence probability when two users have exactly one common neighbor.

Definition 3. Structural Diversity Signature: Given a network $G = \{V, E\}$, its structural diversity signature is defined as a vector of relative link existence rates with respect to the specified common neighborhoods.

Consider, for example, user pairs with between two and four common neighbors. The common neighborhoods represented by these pairs of users correspond to a vector of the relative link existence rates for 17 subgraphs (2 subgraphs for common neighborhoods with size two, 4 for those with size three, and 11 for those with size four).

Given this input and output, our work seeks to understand the underlying driving forces behind link formation and network organization by answering the following questions:

- Does the structural diversity signature vary across networks?
- Is the structural diversity signature a fundamental property of networks?
- How does the structural diversity of common neighborhoods influence the link existence probability?
- Does structural diversity concord or conflict with the principle of homophily in networks?
- Can structural diversity help to improve link inference?

4. STRUCTURAL DIVERSITY SIGNATURE FOR NETWORK SUPERFAMILIES

There exist natural laws that govern the global structure of network systems, through which constant, universal properties such as long-tailed degree distributions arise. However, even when subject to the global properties prescribed by these laws, different networks can still reveal distinct local properties and structures. One striking example is the discovery that networks with long-tailed degree distributions can be naturally catalogued into distinct superfamilies of networks based on their subgraph frequencies [29]. In this section, we investigate how the structural diversity signature can—like subgraph frequency—uncover previously undiscovered mechanisms of network organization, thereby allowing it to serve as a fundamental property by which to catalogue networks.

4.1 Can the Structural Diversity Signature Identify Distinct Network Superfamilies?

To answer this question, we examine the similarity between the functions of structural diversity of common neighborhoods across the 116 networks. We begin by constructing, for each network, the structural diversity signature (a 17-length vector) with two to four common neighbors. We then use this signature to compute the similarity in the structural diversity for each pair of networks.

Figure 4 visualizes the similarity matrix of the structural diversity profiles between every pair of networks. The x - and y -axes represent all 116 networks studied in this work, and the spectrum color represents the correlation coefficient. Note that the arrangement of rows and columns in the presented similarity matrix is determined

by the Ward variance minimization algorithm for hierarchical clustering [42], and the similarity between structural diversity profiles is measured by the Pearson correlation coefficient [29].

Network Superfamilies. We observe four clear, dense clusters in Figure 4 (a). Three of these clusters correspond to real networks, and are labeled as red, blue, and gold at the top dendrogram. The remaining cluster, which corresponds to most of random graphs, is labeled as black. The fact that our similarity analysis distinguishes between real and artificial networks indicates that the structural diversity signatures are able to capture hidden properties underlying the network structures.

According to the structural diversity signature, a total of 38 networks—including Facebook and Friendster—are grouped together into a single cluster (colored ‘red’ in the dendrogram), wherein the structural diversity of common neighborhoods has similar effects on link formation in each network. Another 20 networks—including LinkedIn and BlogCatalog—are grouped into another cluster (colored ‘blue’), indicating strong correlations among these networks but weak correlations between these networks and those in the red cluster. An additional 24 networks are grouped into the final cluster of real networks (colored ‘gold’). While the networks within these three superfamilies demonstrate a high degree of similarity, we note that the networks in the gold cluster demonstrate relatively higher similarity with networks in the red and blue clusters than the networks in the red and blue clusters demonstrate with each other. Finally, there are 34 remaining networks (8 real networks and 26 random graphs) that are not clustered into the three aforementioned clusters (colored ‘black’).

We find that the vast majority of real networks are, based on their structural diversity signatures, clustered into three major superfamilies (i.e., colored red, blue, and gold in the dendrogram). Each superfamily consists of different types of networks, whereas the network type is defined by intuition, according to either the network service or function, such as social (soc-), collaboration (ca-), communication (comm-), location based (loc-) networks, and so on. For example, the ‘red’ superfamily is mainly composed of social (soc-), film and academic collaboration (ca-), and email and mobile communication (comm-) networks. On the other hand, the same type of networks are also grouped into different superfamilies. According to conventional wisdom, for example, the Facebook and LinkedIn networks both belong to the concept of online social networks (soc-). However, the Facebook network is indexed in the ‘red’ superfamily, while the LinkedIn network is indexed in the ‘blue’ superfamily, demonstrating that the structural diversity of common neighborhoods actually serves opposing roles in determining link existence within these two networks.

Random Graphs. We further study how the structural diversity signatures qualify the nature of random graphs. Observed from Figure 4 (a), all Erdős-Rényi (ER) graphs [13] are densely clustered into the bottom left hierarchy. According to Watts and Strogatz [43], WS random graphs with the β parameter close to 1 tend to approach ER random graphs. This theory is captured by the structural diversity profile, as WS graphs with $\beta=0.8$ are successfully grouped into the ER cluster. Further, we find the Barabási-Albert (BA) random graphs cannot be successfully clustered—neither together or nor into other clusters—despite the BA model being able to simulate networks that satisfy the scale-free property [4].

What we find most striking is that the only random graphs that are clustered into the three major clusters (i.e., colored red, blue, or gold) are WS graphs with a small rewiring probability β (e.g., $\beta = 0.2$). A small β in the WS model leads to networks small-world properties, such as short average path lengths and high clus-

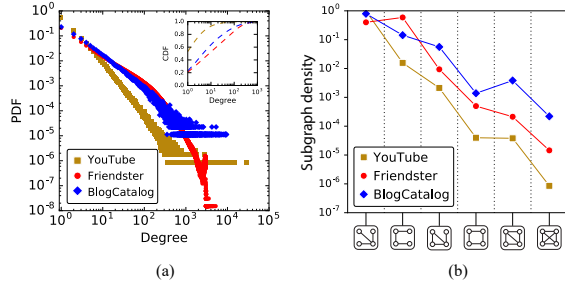


Figure 5: Degree (a) and subgraph frequency (b) distributions.

tering coefficients. According to the similarity matrix, however, the WS model can only imitate one specific superfamily (colored blue) of real networks (e.g., BlogCatalog and LinkedIn). These findings demonstrate that none of the three classical random graph models are able to simulate an important family (red) of real-world networks (which includes networks like Facebook, Friendster, and MySpace), although the random graphs may satisfy a series of network properties, including scale-free and small-world phenomena.

To provide a deeper understanding of the structural diversity of common neighborhoods, we focus on in-depth investigations into the following three large-scale social networks, each of which represent a particular network superfamily: **Friendster** (65,608,366 nodes and 1,806,067,135 edges) from the ‘red’ superfamily, **BlogCatalog** (88,784 nodes and 2,093,195 edges) from the ‘blue’ superfamily, and **YouTube** (1,134,890 nodes and 2,987,624 edges) from the ‘gold’ superfamily. Nevertheless, our findings generally hold for any network within a given superfamily.

4.2 Is the Structural Diversity Signature a Fundamental Network Property?

Characterization of the global similarity and difference across multiple networks is conventionally focused on degree distribution [4, 14], degree sequence [31, 12], and subgraph frequency [29]. To examine the significance of the structural diversity signature, we need to investigate not only its ability to effectively characterize networks, but the extent to which these characterizations are distinct from those provided by conventional methods. Therefore, a crucial question remains: Does the structural diversity signature serve as a general, fundamental property of networks?

To answer this question, we analyze the structural diversity signature at the micro and macro scales over three representative networks (as determined by the superfamilies discovered in Section 4.1)—namely, Friendster, BlogCatalog, and YouTube. As there are several ways to quantify network properties at the global scale, we compare the structural diversity signature with the following four conventional approaches: (1) The subgraph significance profile, a numerical vector of the frequencies (significance level) of different subgraphs [29]. (2) A sequence-of-percentile-degrees vector of node degrees that are ranked at particular positions (e.g., 0%, 10%, 20%, ..., 90%, 100%) of a network’s degree sequence. (3) A bag-of-degrees vector of occurrence counts of node degrees in a network, which is equal to its degree distribution. (4) A bag-of-#CNs vector, in which the occurrence counts of common neighborhood size is vectorized.

At the micro scale, we can examine the visualized distributions of the aforementioned measures. The four-subgraph distributions (computed by ESCAPE [7]) are shown in Figure 5(b), degree distributions in Figure 5(a), and common neighborhood size distribution in Figure 3(b). From these figures, we can observe that each type of distribution reveals similar trends and shapes within the three networks, although they are not identical. We also provide numerical

results of the differences between Friendster and BlogCatalog. The correlation coefficients based on subgraph, sequence of percentile degrees, bag of degrees, and bag of #CNs are 0.579, 1.000, 0.996, and 0.957, respectively, which are significantly higher than structural diversity signature based quantification (−0.267). The strong correlations between Friendster and BlogCatalog produced by all four alternative methods further highlight their inability to uncover hidden network properties.



At the macro scale, we can examine heatmaps of the correlation coefficient matrix for each method. The heatmap for the structural diversity signature is shown in Figure 4(a), while the heatmap for the alternative methods are shown in Figure 4(b)(c)(d)(e). To compare with the structural diversity signature, the ordering of networks in these four matrices is kept identical to that in Figure 4(a). The four resulting matrices fail to show clear and dense clusters, further confirming the unmatched ability of structural diversity signatures to detect unique network superfamilies. Note that the subgraph significance profile (Figure 4(b)) is able to categorize the networks into multiple superfamilies if the same clustering algorithm is applied to the correlation matrix. Based on these results, we argue that the structural diversity signature is able to capture underlying mechanisms of network organization that cannot be discovered by conventional methods such as the subgraph significance profile [29], degree distribution [14], and degree sequence [31].

Conclusion. Our comprehensive study based on both micro- and macro-level phenomena demonstrates that the structural diversity signature can detect intrinsic, hidden network superfamilies that are not discoverable by conventional methods. These findings suggest that the structural diversity signature serves as a unique, fundamental property intrinsic to networks.

5. STRUCTURAL DIVERSITY IN LINK EXISTENCE

By leveraging the structural diversity signature, we discover three major superfamilies from 84 real-world networks. To further understand how the structural diversity of common neighborhoods influences link existence in three superfamilies, we focus our investigations on the following three large-scale social networks, each of which represents a particular network superfamily: Friendster from the ‘red’ superfamily, BlogCatalog from the ‘blue’ superfamily, and YouTube from the ‘gold’ superfamily in Figure 4 (a). *However, these findings generally hold for any network within a given superfamily.*

5.1 How Does Structural Diversity Influence Link Existence?

When we control for the size of common neighborhood between two users, how does its structure influence the probability that they form a link in the network? An illustrative example of this question is introduced as follows. Given that two users v_i and v_j have four common neighbors, are they more likely to connect with each other if their four common neighbors do not know each other () or if their four common neighbors already know each other () , i.e.,

$$P(e_{ij}=1 \mid \text{four isolated nodes}) \gtrless P(e_{ij}=1 \mid \text{four nodes in a square}) ?$$

To address this question, we compute 546 billion, 612 million, and 1.26 billion pairs of users in the Friendster, BlogCatalog, and YouTube networks, respectively. Figure 6 presents the relative link existence rates for two, three, four, five, and six common neighborhoods in the Friendster, BlogCatalog, and YouTube networks. It is immediately observable that the impact of structural diversity on link existence in the three networks is remarkably different.

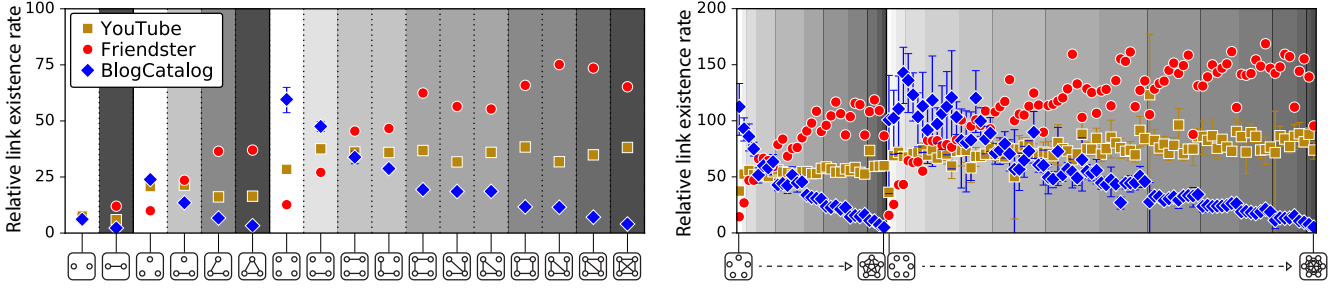


Figure 6: Structural diversity of common neighborhoods in link existence. The three colors of each network—red, blue, and gold—are in accordance with the three superfamily hierarchies of the dendrogram in Figure 4, respectively. x -axis: two-node, three-node, and four-node common neighborhoods on the left side; five-node and six-node common neighborhoods on the right side. The x -axis is ordered according to the following keys: common neighborhood size (ascending), edge density of the common neighborhood (ascending), and component count of the common neighborhood (ascending). When all three keys are the same, the degree sequence of the common neighborhood is in descending order. Shading indicates differences in edge density. Error bars designate the 95% confidence interval.

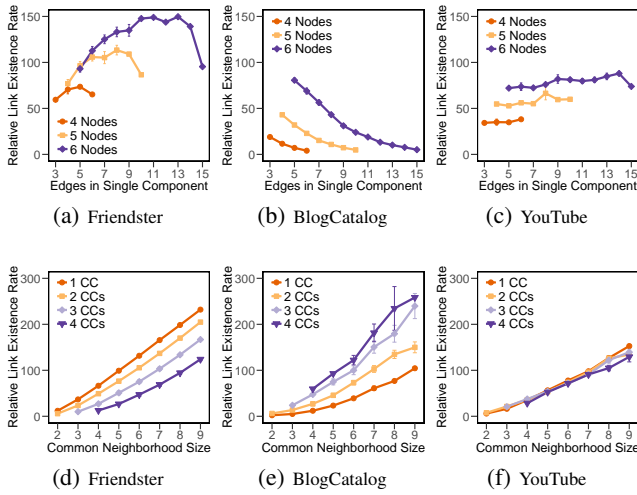









Figure 7: Density and variety vs. link existence. (a)(b)(c) Link existence rate as a function of edge count (density) with one component. (d)(e)(f) Link existence rate as a function of component count (variety).

We study in detail how the structural diversity—that is, the density and variety—of common neighborhoods influences the link existence. Recall that density is measured by the ratio of the number of actual edges to the number of possible edges between all pairs of nodes, and variety is measured by the number of connected components. In general, if we control the size of common neighborhood, then as the component count increases, the link existence rate decreases in Friendster (and its superfamily) but increases in BlogCatalog (and its superfamily). This finding is illustrated in Figure 6 and Figure 7(d)(e). This tells us that users on BlogCatalog are more likely to connect if their common friends are more structurally diverse, while users on Friendster are more likely to connect if their common friends are densely embedded in the same community.

For common neighborhoods with the same size and component count, we still observe variations in the link existence rate. We further examine the impact of edge density among common neighborhoods, as distinguished by different shadings in Figure 6. For a given size of common neighborhood, if we focus on common neighbor subgraphs with the same densities (shadings), the link existence rates are relatively similar to each other. For example, when two users have four common neighbors, they have similar probabil-

Table 1: Correlation analysis for relative link existence rate.

Network	#CN	2	3	4	5	6
Friendster	Density	1.0	0.94	0.89	0.84	0.81
	#Components	-1.0	-0.99	-0.95	-0.88	-0.79
BlogCatalog	Density	-1.0	-0.97	-0.95	-0.95	-0.92
	#Components	1.0	0.98	0.94	0.89	0.67
YouTube	Density	-1.0	-0.84	0.40	0.67	0.50
	#Components	1.0	0.86	-0.38	-0.55	-0.38

ities to connect if their common neighborhood forms the following structures:  or  (four edges),  or  or  (three edges),  or  (two edges). On the other hand, with increasing densities in common neighborhoods, as in Figure 6 and Figure 7(a)(b), the relative link existence rate increases in Friendster but decreases in BlogCatalog. Indeed, the edge density of the common neighborhood also has a crucial role in determining link existence in both Friendster and BlogCatalog networks.

Recall that a common neighborhood with more components and sparser edges is considered more diverse. If we fix the number of common neighbors, we find that the structural diversity of common neighborhoods has a negative effect on the formation of online friendships in Friendster (and its superfamily) but a positive effect in BlogCatalog (and its superfamily). This observation reveals a fundamental difference between these two networks and their superfamilies both in their microscopic structures and link formation mechanisms.

Further, we quantify the impact of density and variety on link existence. Table 1 reports the Pearson correlation coefficients ρ between the relative link existence rate and the structural diversity of common neighborhoods. We can clearly see that both variety (#components) and density are strongly correlated ($|\rho| > 0.8$) with the link existence rate in Friendster and BlogCatalog, although one is positively correlated and the other one is negatively correlated.

Conclusion. We demonstrate that the structural diversity of common neighborhoods as quantified by density and variety is a crucial factor in determining link existence across networks. Further, the contrasting influences of structural diversity on the link existence correspond to networks catalogue to different superfamilies (shown in Figure 4). This observation reaffirms our claim that the structural diversity signature—the means by which we organize these superfamilies—serves as a fundamental property of networks.

5.2 Does Structural Diversity Violate the Principle of Homophily?

Previously, we demonstrated that in the ‘red’ network superfamily (Friendster) the structural diversity of common neighborhoods is in general negatively associated with link existence, i.e., $P(e=1|\text{Ⓢ}) < P(e=1|\text{Ⓣ})$, while in the ‘blue’ superfamily (BlogCatalog) it is in general positively associated with link existence i.e., $P(e=1|\text{Ⓢ}) > P(e=1|\text{Ⓣ})$. A subsequent question that one may ask is whether structural diversity conflicts with the principle of homophily. Specifically, for Friendster, this can be formalized as:

$$P(e=1|\text{Ⓢ}) > P(e=1|\text{Ⓣ})?$$

In BlogCatalog, the question can be similarly formalized as:

$$P(e=1|\text{Ⓢ}) > P(e=1|\text{Ⓣ})?$$

Conventional wisdom may answer “yes” to both cases, as the concept of structural homophily suggests that, all other things equal, relationships are more likely to form between individuals that share a larger common neighborhood. Surprisingly, however, we find that there is no empirical evidence to support the existence of homophily within the context of structural diversity.

In Figure 6, we can observe that the link existence rate between two BlogCatalog users with densely connected common neighbors is actually *lower* than the link existence rate between users with fewer but more loosely connected (less dense) neighbors. For example, if two users share four common neighbors, the probability that there exists a link between them is, in more than half of the eleven configurations ($\text{Ⓢ}, \text{Ⓣ}, \text{Ⓢ}, \text{Ⓣ}, \text{Ⓢ}, \text{Ⓣ}, \text{Ⓢ}$), lower than the probability of a link between users that share three disconnected common neighbors (Ⓢ). In fact, $P(e=1|\text{Ⓢ})$ is 493% higher than $P(e=1|\text{Ⓣ})$; even $P(e=1|\text{Ⓢ})$ is higher than $P(e=1|\text{Ⓣ})$. By contrast, in Friendster, homophily is instead violated when a larger size of disjoint common neighbors meets with a smaller size of connected ones. For example, the link existence probabilities given Ⓢ and Ⓣ are lower than those given Ⓢ and Ⓣ . Similar violations can be seen to occur in various cases with different numbers of common neighborhoods.

Conclusion. Our observations show that structural diversity does indeed, in many cases, violate the principle of homophily. These observations suggest that the fundamental assumption held by the homophily principle that “more common friends means a higher probability to connect” can often be an oversimplification, and—as clearly shown—is not necessarily true.

5.3 Can Structural Diversity Help to Improve Link Inference?

The structural diversity of common neighborhoods is crucial in determining link existence. Often it violates the principle of structural homophily, which demonstrates a simple and yet effective predictor for inferring link existence. Accordingly, we further explore the extent to which structural diversity can help link inference. More formally, we ask which of the following measurements is more accurate,

$$P(e=1|\text{Ⓢ}) \text{ or } P(e=1|\text{#CN}=4)?$$

Note that to answer this question, we focus on qualifying the effect of structural diversity in link inference. Therefore, we leave the improvement of link prediction performance for future work.

First, to demonstrate the role of structural diversity in link inference, we perform a regression analysis, shown in Table 2. We find that, together with the size of common neighborhoods (#CN), the two characteristics of structural diversity—density and variety

Table 2: Regression analysis for link existence rate. Significance code: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Network	Friendster	BlogCatalog	YouTube
Intercept	−0.03845 ***	0.00010	−0.01855 ***
#CN	0.01948 ***	0.00252 ***	0.00792 ***
Density	0.03234 ***	−0.01580 ***	0.00563 **
#Components	−0.01102 ***	0.00114 ***	−0.00047
Adj. R^2 (Diversity)	0.83330	0.76750	0.81440
Adj. R^2 (Homophily)	0.42300	0.14260	0.77160

Table 3: Inferring link existence.

Metric	Method	Friendster	BlogCatalog	YouTube
Data	#Pairs	67,033,108,105	224,786,028	118,635,122
	%Positive	0.9183%	0.0943%	0.5082%
AUPR	Homophily	0.02230	0.00178	0.01524
	Diversity	0.03499	0.00279	0.01532
AUROC	Homophily	0.68539	0.66259	0.69371
	Diversity	0.71722	0.70239	0.68401

(#Components)—can be used as highly accurate predictors of the link existence rate ($R^2 > 0.75$). We also find that the density and variety of common neighborhoods serve as statistically significant ($p < 0.001$) factors in the Friendster and BlogCatalog networks. Observed from the last row of Table 2, when predicting for the Friendster and BlogCatalog networks, we can achieve a far better estimation by using the structural diversity of common neighborhoods than using only structural homophily (#CN), as measured by R^2 . On Friendster, R^2 improves from 0.42 to 0.83 (+97%), and on BlogCatalog, R^2 improves from 0.14 to 0.76 (+442%).

Second, we use both structural homophily and diversity as link predictors to infer whether there exists a link between two users. For structural homophily, we use #CN as the unsupervised predictor. For structural diversity, we use the linear combination of its two characteristics—density and variety—as the predictor. For these predictions, we limit the candidate pairs of users to be inferred as those users with between two and six common neighbors. This generates more than 67 billion, 224 million, and 118 million data instances in the Friendster, BlogCatalog, and YouTube networks, respectively. We further note that the ratio between positive (existing links) and negative (non-existing links) instances is highly imbalanced in each network, resulting in difficult prediction tasks.

Table 3 shows the link inference performance generated by structural homophily and diversity on each of the three networks as measured by AUPR and AUROC. Figure 8 illustrates the corresponding precision-recall curves. In terms of AUPR, the structural diversity-based unsupervised predictor outperforms the homophily-based predictor by about 57% in the Friendster and BlogCatalog networks. In terms of AUROC, structural diversity also demonstrates greater predictive power than homophily. An application of the t -test to these results finds that the improvements of the diversity-based predictor over homophily-based predictor are highly statistically significant ($p \ll 0.001$).

Note that the structural diversity-based predictor does not outperform the structural homophily-based predictor on the YouTube network. While the lack of improvement could be considered disappointing, this result actually further validates the findings in Figure 6, which shows that the impact of the structural diversity of common neighborhoods on link existence can differ among networks of different superfamilies. Specifically, this result shows that

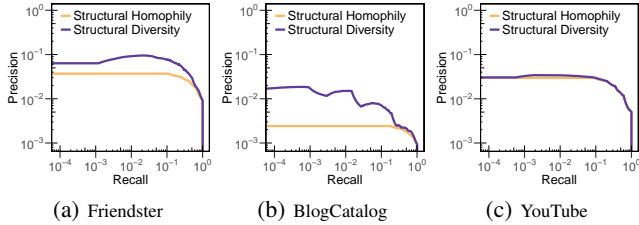


Figure 8: Precision-recall curves for inference of link existence.

the influence of structural diversity on networks in the ‘gold’ superfamily, which includes the YouTube network, is not as significant as it is for the ‘blue’ and ‘red’ superfamilies. That is, the observed difference in performance is a consequence of the underlying factors that distinguish the ‘gold’ superfamily from the others.

Conclusion. We provide empirical evidence that the structural diversity of common neighborhoods helps the link inference task for networks in the ‘blue’ and ‘red’ superfamilies, and we demonstrate that this performance reaffirms the existence of superfamilies. As a result, we find that the proper application of structural diversity has the potential to substantially improve the predictability of link existence, with important implications for improving recommendation functions employed by social networking sites.

6. RELATED WORK

Social theories are the empirical abstraction and interpretation of social phenomena at a societal scale. The idea of homophily, in particular, dates back thousands of years to Aristotle, who observed that people “love those who are like themselves” [1]. In its modern-day conception, the principle of homophily holds that individuals are more likely to associate and bond with similar others [21, 28].

Homophily and Embeddedness. In the context of network science, structural homophily suggests that people with more common neighbors tend to connect with each other [32, 18, 17]. The number of common neighbors is often termed embeddedness [27, 41, 11]. A long line of work has demonstrated the power of structural homophily in determining link existence [25, 17, 3, 9]. Backstrom and Leskovec, for example, showed that the probability that two Facebook users become online friends is exceptionally low if they share only a small number of mutual friends [3]. But even with a fixed number of common neighbors, there still exist a diverse set of structures that describe common neighborhoods. How these diverse structures influence link existence remains an open question in both network and social science.

Structural Diversity. The concept of structural diversity was first proposed by Ugander et al. [39], who found that the user recruitment rate in Facebook is determined by the variety of an individual’s contact neighborhood, rather than the size of his or her neighborhood. Further studies show that the diversity of one’s ego network also has significant influence on a user’s other social decisions, such as whether to pay for online gaming [15]. The difference between this work and our study centers around neighborhood studies. While Ugander et al. study the variety of a single individual’s contact neighborhood, we instead focus on the structural diversity of a pair of individual’s common neighborhoods.

Other studies have leveraged the concept of common neighborhoods. Dong et al. incorporated the edge count of common neighborhoods into a link prediction model [8, 5]. Backstrom and Kleinberg designed a new tie strength metric—dispersion—based on the connections among mutual friends, which can be used to accurately

infer one’s significant other (e.g., husband or wife and fiancé or fiancée) from Facebook. Their goal, however, is to classify the type of existing social tie, while we aim to quantify the diversity of common neighborhoods of both connected and non-connected users.

Subgraph and Network Superfamily. Our work is also related to subgraph (motif) frequency. Milo et al. investigated the distributions of subgraph frequency across multiple types of networks, and proposed a subgraph-based significance profile for networks [29]. By leveraging this profile, they discovered several network superfamilies, whereby networks in the same superfamily display similar subgraph distributions. Recently, Ugander et al. developed a framework to investigate subgraph frequencies in real networks, which is able to characterize both the empirical as well as extremal geography of large graphs [38]. However, our work is different from subgraph mining [29, 38], graph classification [19], and the graph isomorphism problem [40]. Instead of these topics, our focus is on uncovering the principles that drive the formation of local network structure and exploring the significance of structural diversity in driving link organization and network superfamily detection.

Finally, the structural diversity of common neighborhoods also offers substantial potential for applications to other important network mining tasks, including link prediction [25, 10], social recommendation [26, 36], tie strength [16, 2], and network evolution [22, 18, 23]. Structural diversity also has connections with heterogeneous network analysis [35], wherein diversity can be measured by the different types of nodes and links.

7. CONCLUSION

In this paper, we study how the different structural configurations of common neighborhoods can influence the link existence probability, and we examine the implications of these observations for how we organize networks. Through a comprehensive study of 116 large-scale real-world and random networks, we conclude that, controlling for the number of common neighbors, the structure of common neighborhoods—particularly their density and variety—exhibits substantial influence on link existence rates across a wide variety of networks. We also find that the structural diversity of common neighborhoods has a positive influence in some networks (e.g., LinkedIn and BlogCatalog) but a negative influence in others (e.g., Facebook and Friendster), signifying an intrinsic difference in microscopic network structures and link formation mechanisms.

Surprisingly, although the principle of homophily has been acknowledged to hold over a wide variety of networks, we find that structural diversity demonstrates properties that conflict with that of homophily. We demonstrate cases where a pair of users that share four densely connected common neighbors are actually less likely to connect than a pair of users with two or three disconnected common neighbors. We find that because structural diversity captures information like this, it can benefit the task of link inference, improving predictability within our networks by up to 57% as measured by AUPR. Given that a considerable number of link inference methods are based exclusively on the homophily principle, the adoption of structural diversity provides new opportunities for tasks such as “People You May Know” in Facebook.

We further define the structural diversity signature, which serves as a fundamental property of a network similar to degree distribution, degree sequence, and subgraph distribution, but which reveals previously undiscovered mechanisms of network organization. For example, while networks such as Friendster and BlogCatalog show similar degree distributions and subgraph distributions, our structural diversity signature captures the intrinsic differences in link formation mechanisms and classifies them into different network

superfamilies. Furthermore, a study of representative random graph models (ER, WS and BA) shows that only the WS model can imitate a particular superfamily of real-world networks, while none of the models can simulate the other two superfamilies of real-world networks (e.g., Facebook, Friendster, GooglePlus and Phone call). This not only demonstrates the power of the structural diversity signature as a new, fundamental property of networks, but also provides new opportunities for building random graph models.

We acknowledge that not all of the networks demonstrate remarkable differences in the link existence rates for diverse common neighborhoods (e.g., YouTube), but our comprehensive study across 116 networks does show that structural diversity has a pervasive influence on networks from a wide range of domains. By showing that the structural diversity of common neighborhoods can have different influences on link formation rates, we open new pathways to the study of network evolution, link prediction, triadic closure, and so on. Our next step will be to extend our examination of the structural diversity of common neighborhoods beyond homogeneous and static networks, further including heterogeneous networks and dynamic, inter-genre networks. Moreover, we would like to further examine the interrelations between the two characteristics of structural diversity of common neighborhoods—density and variety. Finally, we intend to incorporate the structural diversity signature into machine learning frameworks to improve link prediction and recommendation performance.

Acknowledgments. We would like to thank Xian Wu for discussions. The work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and the National Science Foundation (NSF) Grants BCS-1229450 and IIS-1447795.

8. REFERENCES

- [1] N. E. Aristotle and V. Book. Aristotle in 23 volumes, vol. 19, translated by h. rackham, 1934.
- [2] L. Backstrom and J. M. Kleinberg. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *CSCW '14*, pages 831–841, 2014.
- [3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM '11*, pages 635–644, 2011.
- [4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [5] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports*, 3, 2013.
- [6] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.
- [7] S. Comandur. Escape. <https://bitbucket.org/seshadhri/escape>, 2016.
- [8] Y. Dong, Q. Ke, B. Wang, and B. Wu. Link prediction based on local information. In *ASONAM '11*, pages 382–386, 2011.
- [9] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *IEEE ICDM '12*, pages 181–190, 2012.
- [10] Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang. Coupledip: Link prediction in coupled networks. In *KDD '15*, pages 199–208. ACM, 2015.
- [11] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
- [12] P. L. Erdős, I. Miklós, and Z. Toroczkai. New classes of degree sequences with fast mixing swap Markov chain sampling. *ArXiv e-prints*, Jan. 2016.
- [13] P. Erdős and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [14] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99*, pages 251–262, 1999.
- [15] Z. Fang, X. Zhou, J. Tang, W. Shao, A. Fong, L. Sun, Y. Ding, L. Zhou, and J. Luo. Modeling paying behavior in game social networks. In *CIKM '14*, pages 411–420, 2014.
- [16] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
- [17] J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM '11*, pages 1137–1146, 2011.
- [18] E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *Phys. Rev. E*, 64:046132, 2001.
- [19] X. Kong and P. S. Yu. Semi-supervised feature selection for graph classification. In *KDD '12*, pages 793–802, 2010.
- [20] J. Kunegis. Konect: the koblenz network collection. In *WWW '13 companion*, pages 1343–1350, 2013.
- [21] P. F. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, New York: Van Nostrand, pages 8–66, 1954.
- [22] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08*, pages 462–470, 2008.
- [23] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD '05*, pages 177–187, 2005.
- [24] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [25] D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.
- [26] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM '11*, pages 287–296, 2011.
- [27] P. V. Marsden and K. E. Campbell. Measuring tie strength. *Social forces*, 63(2):482–501, 1984.
- [28] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.
- [29] R. Milo, S. Itzkovitz, N. Kashtan, R. Leviit, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, March 2004.
- [30] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC '07*, pages 29–42, 2007.
- [31] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.
- [32] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, 2001.
- [33] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.
- [34] R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI '15*, pages 4292–4293, 2015.
- [35] Y. Sun and J. Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.
- [36] J. Tang, S. Wu, J. Sun, and H. Su. Cross-domain collaboration recommendation. In *KDD '12*, pages 1285–1293, 2012.
- [37] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD '08*, pages 990–998, 2008.
- [38] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *WWW '13*, pages 1307–1318, 2013.
- [39] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *PNAS*, 109(16):5962–5966, 2012.
- [40] J. R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM*, 23(1):31–42, 1976.
- [41] B. Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly*, pages 35–67, 1997.
- [42] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
- [43] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, pages 440–442, Jun 1998.
- [44] R. Zafarani and H. Liu. Social computing data repository at ASU, 2009.

Table 4: The statistics of real networks.

NetworkName	#nodes	#edges	Avg degree	10%	50%	90%	Max degree	average cc	global cc	diameter	#triangles	#triplets
blog- BlogCatalog 1-ASU	88,784	2,093,195	47.1525	1	5	88	9,444	0.3533	0.0624	9	51,193,389	2,410,854,351
blog-BlogCatalog2-ASU	97,884	1,668,647	34.0944	1	4	48	27,849	0.4921	0.0403	5	40,662,527	2,986,356,565
blog-BlogCatalog3-ASU	10,312	333,983	64.7756	4	21	136	3,992	0.4632	0.0973	5	5,608,664	167,281,662
ca-Actor-ND	374,511	15,014,839	80.1837	8	36	166	3,956	0.7788	0.1867	13	346,728,049	5,225,759,780
ca-AstroPh-Newman	14,845	119,652	16.1202	2	8	41	360	0.6696	0.5937	14	754,159	3,056,896
ca-AstroPh-SNAP	17,903	196,972	22.0044	2	10	55	504	0.6328	0.4032	14	1,350,014	8,694,840
ca-CS-AMiner	1,066,379	4,594,140	8.6163	1	4	18	1,983	0.6496	0.1873	24	10,312,677	154,825,662
ca-CS2004to2008-AMiner	434,357	1,578,255	7.2671	2	4	15	723	0.6684	0.3704	27	3,451,345	24,501,202
ca-CS2006to2010-AMiner	543,452	2,066,258	7.6042	2	4	16	1,082	0.6745	0.2939	25	4,371,426	40,256,430
ca-CS2009to2010-AMiner	315,263	1,059,719	6.7228	2	4	14	652	0.6989	0.4180	23	2,115,016	13,063,018
ca-CS2011to2012-Aminers	347,389	1,229,655	7.0794	2	4	14	793	0.7073	0.4002	29	2,584,467	16,787,160
ca-CondMat-SNAP	21,363	91,286	8.5462	2	5	18	279	0.6417	0.3172	15	171,051	1,446,763
ca-CondMat2003-Newman	27,519	116,181	8.4437	2	5	18	202	0.6546	0.3393	16	228,093	1,788,720
ca-CondMat2005-Newman	36,458	171,735	9.4210	2	5	20	278	0.6566	0.2903	18	374,300	3,493,465
ca-DBLP-NetRep	540,486	15,245,729	56.4149	5	34	135	3,299	0.8019	1.1663	23	444,095,058	698,172,615
ca-DBLP-SNAP	317,080	1,049,866	6.6221	1	4	14	343	0.6324	0.3850	23	2,224,385	15,107,734
ca-GrQc-SNAP	4,158	13,422	6.4560	1	3	15	81	0.5569	1.0829	17	47,779	84,582
ca-HepPh-SNAP	11,204	117,619	20.9959	2	5	47	491	0.6216	1.1768	13	3,357,890	5,202,255
ca-HepTh-SNAP	8,638	24,806	5.7435	1	3	13	65	0.4816	0.3460	18	27,869	213,790
ca-Hollywood-NetRep	1,069,126	56,306,653	105.3320	6	31	212	11,467	0.7664	0.3900	12	4,916,220,615	32,896,279,137
ca-IMDB-NetRep	896,305	3,782,447	8.4401	1	2	22	1,590	0.0001	0.0001	12	4,358	161,800,945
ca-MathSci-NetRep	332,689	820,644	4.9334	1	3	11	496	0.4104	0.1504	24	576,778	10,928,378
cit-CiteSeer-KONECT	365,154	1,721,981	9.4315	1	5	19	1,739	0.1832	0.0513	34	1,350,310	77,658,938
cit-Cora-KONECT	23,166	89,157	7.6972	1	5	16	377	0.2660	0.1268	20	78,791	1,786,074
cit-HepPh-d-SNAP	34,401	420,784	24.4635	3	15	54	846	0.2856	0.1613	14	1,276,859	22,468,237
cit-HepTh-d-SNAP	27,400	352,021	25.6950	3	15	57	2,468	0.3139	0.1299	15	1,478,698	32,665,296
cit-Patents-d-SNAP	3,764,117	16,511,740	8.7732	1	6	19	793	0.0758	0.0703	26	7,514,922	313,229,094
comm-CALL-ND	4,295,638	7,893,769	3.6753	1	3	7	110	0.2179	0.1985	45	2,253,963	31,804,482
comm-EmailEnron-SNAP	33,696	180,811	10.7319	1	3	19	1,383	0.5092	0.0903	13	725,311	23,384,268
comm-EmailEuAll-SNAP	32,430	54,397	3.3547	1	1	3	623	0.1127	0.0273	9	48,992	5,341,634
comm-LinuxKernel-KONECT	10,857	76,317	14.0586	1	2	24	1,927	0.3486	0.1185	13	698,240	16,977,912
comm-Mobile-ND	5,324,963	10,410,903	3.9102	1	3	8	22,224	0.1811	0.0104	36	2,895,897	835,604,293
comm-SMS-ND	2,369,078	3,330,086	2.8113	1	2	6	22,224	0.0669	0.0013	42	326,282	770,920,401
comm-WikiTalk-SNAP	92,117	360,767	7.8328	1	1	11	1,220	0.0589	0.0483	11	836,467	51,083,880
loc-Brightkite-SNAP	56,739	212,945	7.5061	1	2	16	1,134	0.1734	0.1193	18	494,408	11,938,424
loc-Foursquare-NetRep	639,014	3,214,986	10.0623	1	1	19	106,218	0.1080	0.0016	4	21,651,003	39,400,700,856
loc-Foursquare-ASU	639,014	3,214,986	10.0623	1	1	19	106,218	0.1080	0.0016	4	21,651,003	39,400,700,856
loc-Gowalla-SNAP	196,591	950,327	9.6681	1	3	20	14,730	0.2367	0.0239	16	2,273,138	283,580,626
soc-Academia-NetRep	137,969	369,692	5.3591	1	2	12	702	0.1421	0.0806	21	220,641	7,995,452
soc-Advogato-KONECT	2,716	7,773	5.7239	1	3	14	138	0.2233	0.1325	13	5,383	116,510
soc-BuzzNet-ASU	101,163	2,763,066	54.6260	2	14	98	64,289	0.2321	0.0108	5	30,919,848	8,542,533,935
soc-Catster-NetRep	148,826	5,447,464	73.2058	3	22	84	80,634	0.3877	0.0111	10	185,462,078	50,059,386,906
soc-Delicious-ASU	536,108	1,365,961	5.0958	1	1	10	3,216	0.0322	0.0106	14	487,972	137,770,815
soc-Digg-ASU	770,799	5,907,132	15.3273	1	2	16	17,643	0.0881	0.0482	18	62,710,792	3,842,962,151
soc-Dogster-NetRep	426,485	8,543,321	40.0639	2	12	58	46,503	0.1710	0.0144	11	83,499,345	17,303,939,974
soc-Douban-ASU	154,908	327,162	4.2240	1	1	5	287	0.0161	0.0104	9	40,612	11,623,280
soc-Epinions1-d-SNAP	75,877	405,739	10.6947	1	2	18	3,044	0.1378	0.0687	15	1,624,481	69,327,677
soc-Facebook-MPI	63,392	816,886	25.7725	1	11	69	1,098	0.2218	0.1639	15	3,501,534	60,606,675
soc-Facebook1-NetRep	3,097,165	23,667,394	15.2833	1	1	40	4,915	0.0970	0.0493	12	55,606,428	3,330,498,121
soc-Flickr-AMiner	214,424	9,114,421	85.0131	2	23	210	10,486	0.1464	0.0832	10	132,139,697	4,630,544,599
soc-Flickr-ASU	80,513	5,899,882	146.5570	5	46	364	5,706	0.1652	0.2142	6	271,601,126	3,531,448,904
soc-Flickr-MPI	1,624,992	15,476,835	19.0485	1	2	19	27,236	0.1892	0.1212	24	548,646,525	13,028,541,364
soc-Flxiter-ASU	2,523,386	7,918,801	6.2763	1	1	7	1,474	0.0834	0.0138	8	7,897,122	1,711,880,027
soc-Friendster-ASU	5,689,498	14,067,887	4.9452	1	1	5	4,423	0.0502	0.0048	9	8,722,131	5,484,816,732
soc- Friendster -SNAP	65,608,366	1,806,067,135	55.0560	1	9	148	5,214	0.1623	0.0176	37	4,173,724,142	708,133,792,538
soc-GooglePlus-NetRep	78,723	319,999	8.1297	1	2	19	538	0.1982	0.2934	59	1,386,340	12,787,287
soc-Hamsterster-KONECT	2,000	16,098	16.0980	2	9	36	273	0.5401	0.2709	10	52,665	530,614
soc-Hyves-ASU	1,402,673	2,777,419	3.9602	1	1	7	31,883	0.0448	0.0016	10	752,401	1,444,870,827
soc-LastFM-Aminers	135,876	1,685,158	24.8044	1	9	54	3,137	0.1983	0.0946	12	9,097,399	279,291,397
soc-LastFM-ASU	1,191,805	4,519,330	7.5840	1	2	11	5,150	0.0727	0.0131	10	3,946,207	898,270,114
soc-Libimseti-KONECT	34,339	124,722	7.2642	1	2	16	613	0.0224	0.0265	15	54,375	6,103,992
soc-LinkedIn-Aminers	6,725,712	19,360,071	5.7570	2	4	11	869	0.3700	0.2863	32	12,862,009	121,917,817
soc-LiveJournal-Aminers	3,017,282	85,654,975	56.7762	4	17	114	910,088	0.1196	0.0017	8	507,338,233	919,635,317,380
soc-LiveJournal-ASU	2,238,731	12,816,184	11.4495	1	2	21	5,873	0.1270	0.0230	8	28,204,049	3,658,174,479
soc-LiveJournal-MPI	5,189,809	48,688,097	18.7630	1	6	45	15,017	0.2749	0.1352	23	310,784,143	6,586,074,658
soc-LiveJournal-SNAP	3,997,962	34,681,189	17.3494	1	6	42	14,815	0.2843	0.1368	21	177,820,130	3,722,307,805
soc-LiveJournal1-d-SNAP	4,843,953	42,845,684	17.6904	1	5	42	20,333	0.2743	0.1280	20	285,688,896	6,412,296,576
soc-LiveMocha-ASU	104,103	2,193,083	42.1329	2	13	91	2,980	0.0544	0.0142	6	3,361,651	706,231,197
soc-MySpace-Aminers	853,360	5,635,236	13.2072	1	6	26	25,105	0.0433	0.0022	14	1,256,533	1,686,861,075
soc-Orkut-NetRep	2,997,166	106,349,209	70.9665	7	42	152	27,466	0.1700	0.0439	9	524,643,952	35,294,034,217
soc-Orkut-SNAP	3,072,441	117,185,083	76.2814	8	45	162	33,313	0.1666	0.0424	9	627,584,181	43,742,714,028
soc-Pokec-d-SNAP	1,632,803	22,301,964	27.3174	1	13	70	14,854	0.1094	0.0483	14	32,557,458	1,988,401,184
soc-Prospers-d-KONECT	89,171	3,329,970	74.6873	3	34	182	6,515	0.0049	0.0031	8	1,158,669	1,108,949,447
soc-Slashdot0811-d-SNAP	77,360	469,180	12.1298	1	2	25	2,539	0.0555	0.0246	12	551,724	66,861,129
soc-Slashdot0902-d-SNAP	82,168	504,230	12.2731	1	2	25	2,552	0.0603	0.0245	13	602,592	73,175,813
soc-WikiVote-d-SNAP	7,066	100,736	28.5129	1	4	82	1,065	0.1419	0.1369	7	608,389	12,720,410
soc- YouTube -MPI	1,134,890	2,987,624	5.2650	1	1	8	28,754	0.0808	0.0062	24	3,056,386	1,465,313,402
web-BaiduBaik-KONECT	2,107,689	16,996,139	16.1277	1	4	29	97,848	0.1171	0.0025	20	25,206,270	30,809,207,121
web-BerkStan-SNAP	654,782	6,581,871	20.1040	2	8	36	84,230	0.6066	0.0069	208	64,520,617	27,786,200,608
web-Google-SNAP	855,802	4,291,352	10.0288	1	5	20	6,332	0.5190	0.0572	24	13,356,298	686,679,376
web-Hudong-KONECT	1,962,418	14,419,760	14.6959	1	5	23	61,440	0.0783	0.0035	16	21,611,635	18,660,819,412
web-Internet-Newman	22,963	48,436	4.2186	1	2	5	2,390	0.2304	0.0112	11	46,873	12,475,042
web-Stanford-SNAP	255,265	1,941,926	15.2150	2	6	29	38,625	0.6189	0.0086	164	11,277,977	3,907,779,392
web-WWW-ND	325,729	1,090,108	6.6933	1	2	12	10,721	0.2346	0.0931	46	8,910,005	278,151,159